Testing parametric models in linear-directional regression

Eduardo García-Portugués[1,2,3,5] Ingrid Van Keilegom[4]
Rosa M. Crujeiras[3] Wenceslao González-Manteiga[3]

Abstract

This paper presents a goodness-of-fit test for parametric regression models with scalar response and directional predictor, that is, a vector on a sphere of arbitrary dimension. The testing procedure is based on the weighted squared distance between a smooth and a parametric regression estimator, where the smooth regression estimator is obtained by a projected local approach. Asymptotic behavior of the test statistic under the null hypothesis and local alternatives is provided, jointly with a consistent bootstrap algorithm for application in practice. A simulation study illustrates the performance of the test in finite samples. The procedure is applied to test a linear model in text mining.

Keywords: bootstrap calibration, directional data, goodness-of-fit test, local linear regression.

Running title: Testing in linear-directional regression.

[1]Department of Mathematical Sciences. University of Copenhagen (Denmark).
[2]The Bioinformatics Centre, Department of Biology. University of Copenhagen (Denmark).
[3]Department of Statistics and Operations Research. University of Santiago de Compostela (Spain).
[4]Institute of Statistics, Biostatistics and Actuarial Sciences. Université catholique de Louvain (Belgium).
[5]Corresponding author. e-mail: egarcia@math.ku.dk.

1 Introduction

Directional data (data on a general sphere of dimension q) appear in a variety of contexts, the simplest one being provided by observations of angles on a circle (circular data). Directional data is present in wind directions or animal orientation (Mardia and Jupp, 2000) and, recently, it has been considered in higher dimensional settings for text mining (Srivastava and Sahami, 2009). In order to identify a statistical pattern within a certain collection of texts, these objects may be represented by a vector on a sphere where each vector component gives the relative frequency of a certain word. From this vector-space representation, text classification can be performed (Banerjee et al., 2005), but other interesting problems such as popularity prediction could be tackled. For instance, a linear-directional regression model could be used to predict the popularity of articles in news aggregators, quantified by the number of comments or views (Tatar et al., 2012), based on the news contents.

When dealing with directional and linear variables at the same time, the joint behavior could be modeled by considering a flexible density estimator (García-Portugués et al., 2013). Nevertheless, a regression approach may be more useful, allowing at the same time for explaining a relation between the variables and for making predictions. Nonparametric regression estimation methods for linear-directional models have been proposed by different authors. For example, Cheng and Wu (2013) introduced a general local linear regression method on manifolds and, quite recently, Di Marzio et al. (2014) presented a local polynomial method when both the predictor and the response are defined on spheres. Despite the flexibility of these estimators, in terms of interpretation of the results, purely parametric models may be more convenient. In this context, goodness-of-fit tests can be designed, providing a tool for assessing a certain parametric linear-directional regression model.

Goodness-of-fit tests for directional data, or including a directional component in the data generating process, have not been deeply studied. For the density case, Boente et al. (2014) provide a nonparametric goodness-of-fit test for directional densities and similar ideas are used by García-Portugués et al. (2015) for directional-linear densities. Except for the ex-

ploratory tool and lack-of-fit test for linear-circular regression developed by Deschepper et al. (2008) there are no other works in the regression context. The related Euclidean literature is extensive: the reader is referred to Hart (1997) for a comprehensive reference and to Härdle and Mammen (1993) and Alcalá et al. (1999) for the most relevant works for this contribution.

This paper presents a goodness-of-fit test for parametric linear-directional regression models. The test is constructed from a projected local regression estimator (Section 2). The asymptotic distribution of the test statistic, based on a weighted squared distance between the nonparametric and parametric fits, is obtained under a family of local alternatives containing the null hypothesis (Section 3). A bootstrap strategy, proved to be consistent, is proposed for the calibration of the test in practice. The performance of the test is checked for finite samples in a simulation study (Section 4) and the test is applied to assess a constrained linear model for news popularity prediction in text mining (Section 5). An appendix contains the proofs of the main results, whereas technical lemmas and further information on the simulation study and data application are provided as Supporting Information (SI).

2 Nonparametric linear-directional regression

Let $\Omega_q = \left\{\mathbf{x} \in \mathbb{R}^{q+1} : ||\mathbf{x}|| = 1\right\}$ denote the q-sphere in $\mathbb{R}^{q+1}$ and ω_q denote both its associated Lebesgue measure and its surface area, $\omega_q = 2\pi^{\frac{q+1}{2}}/\Gamma\left(\frac{q+1}{2}\right)$. A directional density f satisfies $\int_{\Omega_q} f(\mathbf{x})\,\omega_q(d\mathbf{x}) = 1$. From a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of a random variable (rv) $\mathbf{X}$ with density f, Hall et al. (1987) and Bai et al. (1988) introduced the kernel density estimator

$$\hat{f}_h(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} L_h\left(\mathbf{x}, \mathbf{X}_i\right), \quad L_h(\mathbf{x}, \mathbf{y}) = c_{h,q}(L) L\left(\frac{1 - \mathbf{x}^T\mathbf{y}}{h^2}\right), \quad \mathbf{x} \in \Omega_q, \tag{1}$$

where L is a directional kernel, $h > 0$ is the bandwidth parameter and

$$c_{h,q}(L)^{-1} = \lambda_{h,q}(L) h^q = \lambda_q(L) h^q (1 + o(1)) \tag{2}$$

with $\lambda_{h,q}(L) = \omega_{q-1}\int_0^{2h^{-2}} L(r) r^{\frac{q}{2}-1}(2 - rh^2)^{\frac{q}{2}-1}\,dr$ and $\lambda_q(L) = 2^{\frac{q}{2}-1}\omega_{q-1}\int_0^{\infty} L(r) r^{\frac{q}{2}-1}\,dr$.

Assume that $\mathbf{X}$ is the covariate in the regression model

$$Y = m(\mathbf{X}) + \sigma(\mathbf{X})\varepsilon, \tag{3}$$

where Y is a scalar (response) rv, m is the regression function given by the conditional mean ($m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$), and σ^2 is the conditional variance ($\sigma^2(\mathbf{x}) = \mathbb{V}\mathrm{ar}[Y|\mathbf{X} = \mathbf{x}]$). Errors are collected by ε, a rv such that $\mathbb{E}[\varepsilon|\mathbf{X}] = 0$, $\mathbb{E}[\varepsilon^2|\mathbf{X}] = 1$ and $\mathbb{E}[|\varepsilon|^3|\mathbf{X}]$ and $\mathbb{E}[\varepsilon^4|\mathbf{X}]$ are assumed to be bounded rv's. Both $m, f : \Omega_q \longrightarrow \mathbb{R}$ can be extended from Ω_q to $\mathbb{R}^{q+1} \setminus \{\mathbf{0}\}$ by considering a radial projection. This allows the consideration of easily tractable derivatives and the use of Taylor expansions.

A1. m and f are extended from Ω_q to $\mathbb{R}^{q+1} \setminus \{\mathbf{0}\}$ by $m(\mathbf{x}) \equiv m(\mathbf{x}/||\mathbf{x}||)$ and $f(\mathbf{x}) \equiv f(\mathbf{x}/||\mathbf{x}||)$. m is three times and f is twice continuously differentiable and f is bounded away from zero.

Assumption **A1** guarantees that f and m are uniformly bounded in Ω_q. More importantly, the directional derivative of m (and f) in the direction $\mathbf{x}$ and evaluated at $\mathbf{x}$ is zero, *i.e.*, $\mathbf{x}^T \boldsymbol{\nabla} m(\mathbf{x}) = 0$. This is a key fact on the construction of Taylor expansion of m at $\mathbf{X}_i$:

$$\begin{aligned} m(\mathbf{X}_i) &= m(\mathbf{x}) + \boldsymbol{\nabla} m(\mathbf{x})^T (\mathbf{X}_i - \mathbf{x}) + \mathcal{O}\left(\left|\left|\mathbf{X}_i - \mathbf{x})^2\right|\right|\right) \\ &= m(\mathbf{x}) + \boldsymbol{\nabla} m(\mathbf{x})^T \left(\mathbf{I}_{q+1} - \mathbf{x}\mathbf{x}^T\right)(\mathbf{X}_i - \mathbf{x}) + \mathcal{O}\left(||\mathbf{X}_i - \mathbf{x}||^2\right) \\ &\approx \beta_0 + \boldsymbol{\beta}_1^T \mathbf{B}_{\mathbf{x}}^T (\mathbf{X}_i - \mathbf{x}), \end{aligned}$$

where $\mathbf{B}_{\mathbf{x}} = (\mathbf{b}_1, \ldots, \mathbf{b}_q)_{(q+1)\times q}$ is the *projection matrix* that completes $\mathbf{x}$ to an orthonormal basis $\{\mathbf{x}, \mathbf{b}_1, \ldots, \mathbf{b}_q\}$ of $\mathbb{R}^{q+1}$ and satisfies $\mathbf{B}_{\mathbf{x}}^T \mathbf{B}_{\mathbf{x}} = \mathbf{I}_q$ and $\mathbf{B}_{\mathbf{x}} \mathbf{B}_{\mathbf{x}}^T = \sum_{i=1}^q \mathbf{b}_i \mathbf{b}_i^T = \mathbf{I}_{q+1} - \mathbf{x}\mathbf{x}^T$, with $\mathbf{I}_q$ the identity matrix of dimension q.

With this setting, $\beta_0 \in \mathbb{R}$ captures the constant effect in $m(\mathbf{x})$ while $\boldsymbol{\beta}_1 \in \mathbb{R}^q$ contains the linear effects of the *projected gradient* of m given by $\mathbf{B}_{\mathbf{x}}^T \boldsymbol{\nabla} m(\mathbf{x})$. It should be noted that the dimension of $\boldsymbol{\beta}_1$ is the *adequate* for the q-sphere Ω_q, which would be $q + 1$ if an usual Taylor expansion in $\mathbb{R}^{q+1}$ was performed. The *projected local estimator* at $m(\mathbf{x})$ is obtained as the weighted average of local constant (denoted by $p = 0$) or linear ($p = 1$) fits given by β_0 or $\beta_0 + \boldsymbol{\beta}_1^T \mathbf{B}_{\mathbf{x}}^T (\mathbf{X}_i - \mathbf{x})$, respectively. Given the sample $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ from (3), comprised of independent and identically distributed (iid) rv's in $\Omega_q \times \mathbb{R}$, both fits can be

formulated as the weighted least squares problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{q+1}} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \delta_{p,1} \left(\beta_1, \ldots, \beta_q\right)^T \mathbf{B}_{\mathbf{x}}^T (\mathbf{X}_i - \mathbf{x})\right)^2 L_h(\mathbf{x}, \mathbf{X}_i),$$

where $\delta_{r,s}$ is the Kronecker Delta. The solution to the minimization problem is given by

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{\mathcal{X}}_{\mathbf{x},p}^T \boldsymbol{\mathcal{W}}_{\mathbf{x}} \boldsymbol{\mathcal{X}}_{\mathbf{x},p}\right)^{-1} \boldsymbol{\mathcal{X}}_{\mathbf{x},p}^T \boldsymbol{\mathcal{W}}_{\mathbf{x}} \mathbf{Y}, \tag{4}$$

where $\mathbf{Y}$ is the vector of observed responses, $\boldsymbol{\mathcal{W}}_{\mathbf{x}}$ is the diagonal weight matrix with i-th entry $L_h(\mathbf{x}, \mathbf{X}_i)$, $\boldsymbol{\mathcal{X}}_{\mathbf{x},1}$ is the design matrix with i-th row $(1, (\mathbf{X}_i - \mathbf{x})^T \mathbf{B}_{\mathbf{x}})$ and $\boldsymbol{\mathcal{X}}_{\mathbf{x},0} = \mathbf{1}$ ($\mathbf{1}$ stands for a vector of ones whose dimension is determined by the context). The projected local estimator at $\mathbf{x}$ is given by $\hat{\beta}_0 = \hat{m}_{h,p}(\mathbf{x})$ and is a weighted linear combination of the responses ($\mathbf{e}_1$ is a null vector with one in the first component):

$$\hat{m}_{h,p}(\mathbf{x}) = \mathbf{e}_1^T \left(\boldsymbol{\mathcal{X}}_{\mathbf{x},p}^T \boldsymbol{\mathcal{W}}_{\mathbf{x}} \boldsymbol{\mathcal{X}}_{\mathbf{x},p}\right)^{-1} \boldsymbol{\mathcal{X}}_{\mathbf{x},p}^T \boldsymbol{\mathcal{W}}_{\mathbf{x}} \mathbf{Y} = \sum_{i=1}^{n} W_n^p \left(\mathbf{x}, \mathbf{X}_i\right) Y_i, \tag{5}$$

The next assumptions ensure that $\hat{m}_{h,p}$ is a consistent estimator of m:

A2. The conditional variance σ^2 is uniformly continuous and bounded away from zero.

A3. $L : [0, \infty) \to [0, \infty)$ is a continuous and bounded function with exponential decay.

A4. The sequence of bandwidths $h = h_n$ is positive and satisfies $h \to 0$ and $nh^q \to \infty$.

Assumptions **A2** and **A4** are usual assumptions for the multivariate local linear estimator (Ruppert and Wand, 1994). **A3** allows for the use of non-compactly supported kernels, such as the popular *von Mises kernel* $L(r) = e^{-r}$.

Remark 1. *The proposal of Di Marzio et al. (2014) for a local linear estimator of m is rooted on a Taylor expansion of the* sin *and* cos *functions of the tangent-normal decomposition. This leads to an overparametrized design matrix of $q+2$ columns which makes $\boldsymbol{\mathcal{X}}_{\mathbf{x},p}^T \boldsymbol{\mathcal{W}}_{\mathbf{x}} \boldsymbol{\mathcal{X}}_{\mathbf{x},p}$ exactly singular, a fact handled by the authors with a pseudo-inverse. It should be noted that Di Marzio et al. (2014)'s proposal and (5) present some remarkable differences: for the circular case, (5) corresponds to Di Marzio et al. (2009)'s proposal (with parametrization $\kappa \equiv 1/h^2$), but Di Marzio et al. (2014) differs from the aforementioned reference. Although both estimators share the same asymptotics, (5) somehow offers a simpler construction and a more natural connection with previous proposals.*

3 Goodness-of-fit test for linear-directional regression

Assuming that model (3) holds, the goal is to test if the regression function m belongs to the parametric class of functions $\mathcal{M}_\Theta = \{m_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^s\}$. This is equivalent to testing

$$H_0 : m(\mathbf{x}) = m_{\boldsymbol{\theta}_0}(\mathbf{x}), \text{ for all } \mathbf{x} \in \Omega_q, \text{ versus } H_1 : m(\mathbf{x}) \neq m_{\boldsymbol{\theta}_0}(\mathbf{x}), \text{ for some } \mathbf{x} \in \Omega_q,$$

with $\boldsymbol{\theta}_0 \in \Theta$ known (simple hypothesis) or unknown (composite) and where *for all* holds except for a set of probability zero and *for some* holds for a set of positive probability.

The proposed statistic to test H_0 compares the nonparametric estimator with a smoothed parametric estimator in $\mathcal{M}_\Theta$ through a squared weighted norm:

$$T_n = \int_{\Omega_q} \left(\hat{m}_{h,p}(\mathbf{x}) - \mathcal{L}_{h,p} m_{\hat{\boldsymbol{\theta}}}(\mathbf{x})\right)^2 \hat{f}_h(\mathbf{x}) w(\mathbf{x})\, \omega_q(d\mathbf{x})$$

where $\mathcal{L}_{h,p} m(\mathbf{x}) = \sum_{i=1}^n W_n^p(\mathbf{x}, \mathbf{X}_i)\, m(\mathbf{X}_i)$ represents the local smoothing of the function m from measurements $\{\mathbf{X}_i\}_{i=1}^n$ and $\hat{\boldsymbol{\theta}}$ denotes either the known parameter $\boldsymbol{\theta}_0$ (simple hypothesis) or a consistent estimator (composite hypothesis; see **A6** below). An equivalent expression for T_n, useful for computational implementation, is $T_n = \int_{\Omega_q} \left(\sum_{i=1}^n W_n^p(\mathbf{x}, \mathbf{X}_i)(Y_i - m_{\hat{\boldsymbol{\theta}}}(\mathbf{X}_i))\right)^2 \hat{f}_h(\mathbf{x}) w(\mathbf{x})\, \omega_q(d\mathbf{x})$. This smoothing of the (possibly estimated) parametric regression function is included to reduce the asymptotic bias (Härdle and Mammen, 1993). Besides, in order to mitigate the effect of the difference between $\hat{m}_{h,p}$ and $m_{\hat{\boldsymbol{\theta}}}$ in sparse areas of the covariate, the squared difference is weighted by a kernel density estimate of $\mathbf{X}$, namely $\hat{f}_h$. In addition, by the inclusion of $\hat{f}_h$, the effects of the unknown density both on the asymptotic bias and variance are removed. Optionally, a weight function $w : \Omega_q \longrightarrow [0, \infty)$ can be considered, for example, to restrict the test to specific regions of Ω_q by an indicator function.

The limit distributions of T_n are analyzed under a family of local alternatives that contains H_0 as a particular case and is asymptotically close to H_0:

$$H_{1P} : m(\mathbf{x}) = m_{\boldsymbol{\theta}_0}(\mathbf{x}) + c_n g(\mathbf{x}), \text{ for all } \mathbf{x} \in \Omega_q,$$

where $m_{\boldsymbol{\theta}_0} \in \mathcal{M}_\Theta$, $g : \Omega_q \longrightarrow \mathbb{R}$ and c_n is a positive sequence such that $c_n \to 0$, for instance $c_n = \left(nh^{\frac{q}{2}}\right)^{-\frac{1}{2}}$. With this framework, H_{1P} becomes H_0 when g is such that $m_{\boldsymbol{\theta}_0} + c_n^{-\frac{1}{2}} g \in \mathcal{M}_\Theta$

($g \equiv 0$, for example) and H_1 when the previous statement does not hold for a set of positive probability. The following regularity conditions on the parametric estimation are required:

A5. $m_{\boldsymbol{\theta}}$ is continuously differentiable as a function of $\boldsymbol{\theta}$, and this derivative is also continuous for $\mathbf{x} \in \Omega_q$.

A6. Under H_0, there exists an $\sqrt{n}$-consistent estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$, *i.e.* $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = \mathcal{O}_{\mathbb{P}}\big(n^{-\frac{1}{2}}\big)$ and such that, under H_1, $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_1 = \mathcal{O}_{\mathbb{P}}\big(n^{-\frac{1}{2}}\big)$ for a certain $\boldsymbol{\theta}_1$.

A7. The function g is continuous.

A8. Under H_{1P} , the $\sqrt{n}$-consistent estimator $\hat{\boldsymbol{\theta}}$ also satisfies $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = \mathcal{O}_{\mathbb{P}}\big(n^{-\frac{1}{2}}\big)$.

Theorem 1 (Limit distributions of T_n)**.** *Under H_{1P}, **A1**–**A6** and **A7**–**A8** if $g \not\equiv 0$,*

$$nh^{\frac{q}{2}} \left(T_n - \frac{\lambda_q(L^2)\lambda_q(L)^{-2}}{nh^q} \int_{\Omega_q} \sigma^2_{\boldsymbol{\theta}_0}(\mathbf{x}) w(\mathbf{x})\, \omega_q(d\mathbf{x}) \right)$$

$$\xrightarrow{d} \begin{cases} \infty, & c_n^2 nh^{\frac{q}{2}} \to \infty, \\ \mathcal{N}\left(\int_{\Omega_q} g(\mathbf{x})^2 f(\mathbf{x}) w(\mathbf{x})\, \omega_q(d\mathbf{x}), 2\nu^2_{\boldsymbol{\theta}_0} \right), & c_n^2 nh^{\frac{q}{2}} \to \delta,\ 0 < \delta < \infty, \\ \mathcal{N}(0, 2\nu^2_{\boldsymbol{\theta}_0}), & c_n^2 nh^{\frac{q}{2}} \to 0, \end{cases}$$

where $\sigma^2_{\boldsymbol{\theta}_0}(\mathbf{x}) = \mathbb{E}\left[(Y - m_{\boldsymbol{\theta}_0}(\mathbf{X}))^2 | \mathbf{X} = \mathbf{x}\right]$ is the conditional variance under H_0 and

$$\nu^2_{\boldsymbol{\theta}_0} = \int_{\Omega_q} \sigma^4_{\boldsymbol{\theta}_0}(\mathbf{x}) w(\mathbf{x})^2\, \omega_q(d\mathbf{x}) \times \gamma_q \lambda_q(L)^{-4} \int_0^\infty r^{\frac{q}{2}-1} \left\{ \int_0^\infty \rho^{\frac{q}{2}-1} L(\rho) \varphi_q(r,\rho)\, d\rho \right\}^2 dr,$$

$$\varphi_q(r,\rho) = \begin{cases} L\left(r + \rho - 2(r\rho)^{\frac{1}{2}}\right) + L\left(r + \rho + 2(r\rho)^{\frac{1}{2}}\right), & q = 1, \\ \int_{-1}^1 (1-\theta^2)^{\frac{q-3}{2}} L\left(r + \rho - 2\theta(r\rho)^{\frac{1}{2}}\right) d\theta, & q \geq 2, \end{cases}$$

$$\gamma_q = \begin{cases} 2^{-\frac{1}{2}}, & q = 1, \\ \omega_{q-1}\omega_{q-2}^2 2^{\frac{3q}{2}-3}, & q \geq 2. \end{cases}$$

The convergence rate as well as the asymptotic bias and variance agree with the results in the Euclidean setting given by Härdle and Mammen (1993) and Alcalá et al. (1999), except for the cancellation of the design density in the bias and variance, achieved by the inclusion of $\hat{f}_h$ in the test statistic. The use of a local estimator with $p = 0$ or $p = 1$ does not affect the limiting distribution, given that the *equivalent kernel* (Fan and Gijbels, 1996) is the same (as

seen in the SI). Finally, the general complex structure of the asymptotic bias and variance turns much simpler with the von Mises kernel:

$$\nu^2 = \int_{\Omega_q} \sigma^4(\mathbf{x}) w(\mathbf{x})^2 \, \omega_q(d\mathbf{x}) \times (8\pi)^{-\frac{q}{2}}, \quad \lambda_q(L^2)\lambda_q(L)^{-2} = \left(2\pi^{\frac{1}{2}}\right)^{-q}.$$

3.1 Bootstrap calibration

The distribution of T_n under H_0 can be approximated by the one of its bootstrapped version T_n^*, which can be arbitrarily well approximated by Monte Carlo by generating bootstrap samples. Under H_0, the bootstrap responses are obtained from the parametric fit and bootstrap errors that imitate the conditional variance by a wild bootstrap procedure: $Y_i^* = m_{\hat{\boldsymbol{\theta}}}(\mathbf{X}_i) + \hat{\varepsilon}_i V_i^*$, where $\hat{\varepsilon}_i = Y_i - m_{\hat{\boldsymbol{\theta}}}(\mathbf{X}_i)$ and the variables $V_1^*, \ldots, V_n^*$ are independent from the observed sample and iid with $\mathbb{E}[V_i^*] = 0$, $\mathbb{V}\mathrm{ar}[V_i^*] = 1$ and finite third and fourth moments. A common choice is considering a binary variable with $\mathbb{P}\{V_i^* = (1-\sqrt{5})/2\} = (5+\sqrt{5})/10$ and $\mathbb{P}\{V_i^* = (1+\sqrt{5})/2\} = (5-\sqrt{5})/10$, which corresponds to the *golden section* bootstrap. The test in practice for the composite hypothesis is summarized in the next algorithm (if the simple is considered, set $\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^*$).

Algorithm 1 (Test in practice)**.** *Consider $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ a sample from (3). To test H_0, set a bandwidth h and (optionally) a weight function w and proceed as follows:*

i. Compute $\hat{\boldsymbol{\theta}}$, $\hat{\varepsilon}_i = Y_i - m_{\hat{\boldsymbol{\theta}}}(\mathbf{X}_i)$ and $T_n = \int_{\Omega_q} \left(\sum_{i=1}^n W_n^p(\mathbf{x}, \mathbf{X}_i)\hat{\varepsilon}_i\right)^2 \hat{f}_h(\mathbf{x}) w(\mathbf{x}) \, \omega_q(d\mathbf{x})$.

ii. Bootstrap resampling. For $b = 1, \ldots, B$:

(a) Obtain $\{(\mathbf{X}_i, Y_i^*)\}_{i=1}^n$, where $Y_i^* = m_{\hat{\boldsymbol{\theta}}}(\mathbf{X}_i) + \hat{\varepsilon}_i V_i^*$ and compute $\hat{\boldsymbol{\theta}}^*$ as in i.

(b) Compute $T_n^{*b} = \int_{\Omega_q} \left(\sum_{i=1}^n W_n^p(\mathbf{x}, \mathbf{X}_i)\hat{\varepsilon}_i^*\right)^2 \hat{f}_h(\mathbf{x}) w(\mathbf{x}) \, \omega_q(d\mathbf{x})$ with $\hat{\varepsilon}_i{}^* = Y_i^* - m_{\hat{\boldsymbol{\theta}}^*}(\mathbf{X}_i)$.

iii. Approximate the p-value by $\frac{1}{B}\sum_{b=1}^B \mathbf{1}_{\{T_n \leq T_n^{*b}\}}$.

In order to prove the consistency of the resampling mechanism, that is, that T_n^* has the same asymptotic distribution of T_n, a bootstrap analogue of assumption **A6** is required:

A9. The estimator $\hat{\boldsymbol{\theta}}^*$ computed from $\{(\mathbf{X}_i, Y_i^*)\}_{i=1}^n$ is such that $\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}} = \mathcal{O}_{\mathbb{P}^*}\left(n^{-\frac{1}{2}}\right)$, where $\mathbb{P}^*$ is the probability law conditional on $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$.

From this assumption and Theorem 1 it follows that the probability distribution function (pdf) of T_n^*, conditionally on the sample, converges always in probability to a Gaussian pdf, which is the same asymptotic pdf of T_n if H_0 holds.

Theorem 2 (Bootstrap consistency)**.** *Under **A1**–**A6** and **A9** and conditionally on* $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$,

$$nh^{\frac{q}{2}}\left(T_n^* - \frac{\lambda_q(L^2)\lambda_q(L)^{-2}}{nh^q}\int_{\Omega_q}\sigma^2_{\boldsymbol{\theta}_1}(\mathbf{x})w(\mathbf{x})\,\omega_q(d\mathbf{x})\right) \xrightarrow{d} \mathcal{N}\left(0, 2\nu^2_{\boldsymbol{\theta}_1}\right)$$

in probability. If the null hypothesis holds, then $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0$.

4 Simulation study

The finite sample performance of the goodness-of-fit test is explored in four simulation scenarios, labeled S1 to S4. Their associated parametric regression models are shown in Figure 1 with the following codification: the radius from the origin represents the response $m(\mathbf{x})$ for an $\mathbf{x}$ direction, resulting in a distortion from a perfect circle or sphere. The design densities of the scenarios are taken from García-Portugués (2013), the noise is either heteroskedastic (S1 and S2) or homocedastic (S3 and S4) and two different deviations (for S1–S2 and for S3–S4) are considered. The tests based on the projected local constant and linear estimators are compared with $M = 1000$ Monte Carlo trials and $B = 1000$ bootstrap replicates, under H_0 and H_1, for a grid of bandwidths and with $n = 100$ and $q = 1, 2, 3$. Parametric estimation is done by nonlinear least squares, which is justified by their simplicity and asymptotic normality (Jennrich, 1969), hence satisfying **A6**. For the sake of brevity, only a coarse grained description of the scenarios and a selected output of the study is provided here. The reader is referred to the SI for the complete report.

[Figure 1 around here]

The empirical sizes of the goodness-of-fit tests are shown using the so called *significance trace* (Bowman and Azzalini, 1997), *i.e.*, the curve of percentages of empirical rejections for different bandwidths. As shown in Figure 2, except for very small bandwidths that result in a conservative test, the significance level is stabilized around the 95% confidence band

for the nominal level $\alpha = 0.05$, for the different scenarios and dimensions. The power is satisfactory, given that the proposed tests succeed in detecting the mild deviations from the null hypotheses. Despite the fact that the test based on the local linear estimator ($p = 1$) provides a better power for large bandwidths in certain scenarios, the overall impression is that the test with $p = 0$ is hard to beat: the powers with $p = 0$ and $p = 1$ are almost the same for low dimensions, whereas as the dimension increases the local constant estimator performs better for a wider range of bandwidths. This effect could be explained by the spikes that local linear regression tends to show in the boundaries of the support (design densities of S3 and S4), which become more important as the dimension increases. The lower power for S1 and S4 is due to deviations happening in areas with low density or high variance.

[Figure 2 around here]

5 Application to text mining

In different applications within text mining, it is quite common to consider a *corpus* (collection of documents) and to determine the so-called vector space model: a corpus $\mathbf{d}_1, \ldots, \mathbf{d}_n$ is codified by the set of vectors $\{(d_{i1}, \ldots, d_{iD})\}_{i=1}^n$ (the *document-term matrix*) with respect to a dictionary (or a *bag of words*) $\{w_1, \ldots, w_D\}$, such that d_{ij} represents the frequency of the dictionary's j-th word in the document $\mathbf{d}_i$. Usually, a normalization of the document-term matrix is performed to remove length distortions and map documents with similar contents, albeit different lengths, into close vectors. If the Euclidean norm is used for this, then the documents can then be regarded as points in Ω_{D-1} providing a set of directional data.

The corpus that is analyzed in this application was acquired from the news aggregator *Slashdot* (`wwww.slashdot.org`). This website publishes summaries of news about technology and science that are submitted and evaluated by users. Each news entry includes a title, a summary with links to other related news and a discussion thread gathering users comments. The goal is to test a linear model that takes as a predictor the topic of the news (a directional variable in Ω_{D-1}) and as a response the log-number of comments. This is motivated by the frequent use of simple linear models in this context (see Tatar et al. (2012) for example) and

that, in text classifications, it has been checked that non-linear classifiers hardly provide any advantage with respect to linear ones (Joachims, 2002). After a data preprocessing process (using Meyer et al. (2008); see SI), the $n = 8121$ news collected from 2013 were represented in a document term matrix formed by $D = 1508$ words.

In order to construct a plausible linear model, a preliminary variable selection was performed using LASSO regression with (tuning) parameter λ selected by an overpenalized *three* standard error rule (Hastie et al., 2009). After removing some extra variables by using a backward stepwise method with BIC, a fitted vector $\hat{\boldsymbol{\eta}} \in \mathbb{R}^D$ with $d = 77$ non-zero entries is obtained. The test is applied to check the null hypothesis of a candidate linear model with coefficient $\boldsymbol{\eta}$ constrained to be zero except in these previously selected d words, that is $H_0 : m(\mathbf{x}) = c + \boldsymbol{\eta}^T\mathbf{x}$, with $\boldsymbol{\eta}$ subject to $\mathbf{A}\boldsymbol{\eta} = \mathbf{0}$ for an adequate choice of the matrix $\mathbf{A}_{(D-d)\times D}$. The significance trace of the test (with $p = 0$; $p = 1$ was not implemented due to its higher cost and to computational limitations) presents a minimum p-value of 0.12, hence showing no evidence to reject the linear model for a wide grid of bandwidths. Figure 3 displays a graphical summary of the fitted linear model. As it can be seen, news where stemmed words like "kill", "climat", "polit" appear have a strong positive impact on the number of comments, since these news are likely more controversial and generate broader discussions. On the other hand, scientific related words like "mission", "abstract" or "lab" have a negative impact, as they tend to raise more objective and higher specific discussions. Experiments were conducted with a model of $d = 50$ non-zero coefficients chosen with a higher overpenalization, showing a strong rejection of the null hypothesis.

[Figure 3 around here]

Acknowledgments

We thank professors David E. Losada for his guidance in the data application and Irène Gijbels for her useful theoretical comments. This research was supported by Project MTM2008–03010 from the Spanish Ministry of Science and Innovation, Project 10MDS207015PR from

Dirección Xeral de I+D of the Xunta de Galicia, by IAP research network grant nr. P7/06 of the Belgian government (Belgian Science Policy), by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013) / ERC Grant agreement No. 203650, and by the contract "Projet d'Actions de Recherche Concertées" (ARC) 11/16–039 of the "Communauté française de Belgique" (granted by the "Académie universitaire Louvain"). Work of the first author has been supported by a grant from Fundación Barrié and FPU grant AP2010–0957 from the Spanish Ministry of Education. Authors gratefully acknowledge the computational resources used at the CESGA Supercomputing Center and valuable suggestions by three anonymous referees.

Supporting Information

Supporting information available online contains the technical lemmas used and further information on the simulation study and text mining application.

References

Alcalá, J. T., Cristóbal, J. A., and González-Manteiga, W. (1999). Goodness-of-fit test for linear models based on local polynomials. *Statist. Probab. Lett.*, 42(1):39–46.

Bai, Z. D., Rao, C. R., and Zhao, L. C. (1988). Kernel estimators of density function of directional data. *J. Multivariate Anal.*, 27(1):24–39.

Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S. (2005). Clustering on the unit hypersphere using von Mises-Fisher distributions. *J. Mach. Learn. Res.*, 6:1345–1382.

Boente, G., Rodríguez, D., and González-Manteiga, W. (2014). Goodness-of-fit test for directional data. *Scand. J. Stat.*, 41(1):259–275.

Bowman, A. W. and Azzalini, A. (1997). *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*. Oxford Statistical Science Series. Clarendon Press, Oxford.

Cheng, M.-Y. and Wu, H.-T. (2013). Local linear regression on manifolds and its geometric interpretation. *J. Amer. Statist. Assoc.*, 108(504):1421–1434.

de Jong, P. (1987). A central limit theorem for generalized quadratic forms. *Probab. Theory Related Fields*, 75(2):261–277.

Deschepper, E., Thas, O., and Ottoy, J. P. (2008). Tests and diagnostic plots for detecting lack-of-fit for circular-linear regression models. *Biometrics*, 64(3):912–920.

Di Marzio, M., Panzera, A., and Taylor, C. C. (2009). Local polynomial regression for circular predictors. *Statist. Probab. Lett.*, 79(19):2066–2075.

Di Marzio, M., Panzera, A., and Taylor, C. C. (2014). Nonparametric regression for spherical data. *J. Amer. Statist. Assoc.*, 109(506):748–763.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

García-Portugués, E. (2013). Exact risk improvement of bandwidth selectors for kernel density estimation with directional data. *Electron. J. Stat.*, 7:1655–1685.

García-Portugués, E., Crujeiras, R. M., and González-Manteiga, W. (2013). Kernel density estimation for directional-linear data. *J. Multivariate Anal.*, 121:152–175.

García-Portugués, E., Crujeiras, R. M., and González-Manteiga, W. (2015). Central limit theorems for directional and linear data with applications. *Statist. Sinica*, 25:1207–1229.

Hall, P., Watson, G. S., and Cabrera, J. (1987). Kernel density estimation with spherical data. *Biometrika*, 74(4):751–762.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, 21(4):1926–1947.

Hart, J. D. (1997). *Nonparametric smoothing and lack-of-fit tests*. Springer Series in Statistics. Springer-Verlag, New York.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.*, 40(2):633–643.

Joachims, T. (2002). *Learning to classify text using support vector machines: Methods, theory and algorithms*, volume 668 of *Kluwer International Series in Engineering and Computer Science*. Kluwer Academic Publishers, Boston.

Mardia, K. V. and Jupp, P. E. (2000). *Directional statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester, second edition.

Meyer, D., Hornik, K., and Feinerer, I. (2008). Text mining infrastructure in R. *J. Stat. Softw.*, 25(5):1–54.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.*, 22(3):1346–1370.

Srivastava, A. N. and Sahami, M., editors (2009). *Text mining: classification, clustering, and applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press, Boca Raton.

Tatar, A., Antoniadis, P., De Amorim, M. D., and Fdida, S. (2012). Ranking news articles based on popularity prediction. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 106–110. IEEE.

A Proofs of the main results

Proof of Theorem 1. The proof follows the steps of Härdle and Mammen (1993) and Alcalá et al. (1999). T_n can be separated into three addends by adding and subtracting the true smoothed regression function $T_n = (T_{n,1} + T_{n,2} - 2T_{n,3})(1 + o_{\mathbb{P}}(1))$, where

$$T_{n,1} = \int_{\Omega_q} \left(\sum_{i=1}^{n} W_n^p(\mathbf{x}, \mathbf{X}_i) \left(Y_i - m_{\boldsymbol{\theta}_0}(\mathbf{X}_i) \right) \right)^2 f(\mathbf{x}) w(\mathbf{x}) \, \omega_q(d\mathbf{x}),$$

$$T_{n,2} = \int_{\Omega_q} \left(\mathcal{L}_{h,p}\left(m_{\boldsymbol{\theta}_0} - m_{\hat{\boldsymbol{\theta}}}\right)(\mathbf{x})\right)^2 f(\mathbf{x}) w(\mathbf{x})\, \omega_q(d\mathbf{x}),$$
$$T_{n,3} = \int_{\Omega_q} \left(\hat{m}_{h,p}(\mathbf{x}) - \mathcal{L}_{h,p} m_{\boldsymbol{\theta}_0}(\mathbf{x})\right) \mathcal{L}_{h,p}\left(m_{\boldsymbol{\theta}_0} - m_{\hat{\boldsymbol{\theta}}}\right)(\mathbf{x}) f(\mathbf{x}) w(\mathbf{x})\, \omega_q(d\mathbf{x}),$$

because i of Lemma 4. The proof is divided into the analysis of each addend.

Terms $T_{n,2}$ and $T_{n,3}$. By a Taylor expansion on $m_{\boldsymbol{\theta}}(\mathbf{x})$ as a function of $\boldsymbol{\theta}$ (see **A5**),

$$T_{n,2} = \int_{\Omega_q} \left(\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^T \mathcal{L}_{h,p}\left(\mathcal{O}_{\mathbb{P}}(\mathbf{1})\right)(\mathbf{x})\right)^2 f(\mathbf{x}) w(\mathbf{x})\, \omega_q(d\mathbf{x}) = \mathcal{O}_{\mathbb{P}}\left(n^{-1}\right),$$

because of the boundedness of $\frac{\partial m_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \boldsymbol{\theta}}$ for $\mathbf{x} \in \Omega_q$, **A6** and **A8**. On the other hand,

$$T_{n,3} = \mathcal{O}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right) \int_{\Omega_q} \left(\hat{m}_{h,p}(\mathbf{x}) - \mathcal{L}_{h,p} m_{\boldsymbol{\theta}_0}(\mathbf{x})\right) f(\mathbf{x}) w(\mathbf{x})\, \omega_q(d\mathbf{x}) = \mathcal{O}_{\mathbb{P}}\left(n^{-1}\right),$$

because of the previous considerations and i from Lemma 6. As a consequence, by **A3** it happens that $nh^{\frac{q}{2}} T_{n,3} \xrightarrow{p} 0$ and $nh^{\frac{q}{2}} T_{n,2} \xrightarrow{p} 0$.

Term $T_{n,1}$. $T_{n,1}$ is dealt with $\tilde{L}_h(\mathbf{x}, \mathbf{X}_i) = \frac{1}{nh^q \lambda_q(L) f(\mathbf{x})} L\left(\frac{1 - \mathbf{x}^T \mathbf{X}_i}{h^2}\right)$ from Lemma 5:

$$T_{n,1} = \int_{\Omega_q} \left(\sum_{i=1}^n \tilde{L}_h(\mathbf{x}, \mathbf{X}_i)\left(1 + \boldsymbol{o}_{\mathbb{P}}(1)\right)\left(Y_i - m_{\boldsymbol{\theta}_0}(\mathbf{X}_i)\right)\right)^2 f(\mathbf{x}) w(\mathbf{x})\, \omega_q(d\mathbf{x}) = \widetilde{T}_{n,1}\left(1 + \boldsymbol{o}_{\mathbb{P}}(1)\right).$$

Now it is possible to split $\widetilde{T}_{n,1} = \widetilde{T}_{n,1}^{(1)} + \widetilde{T}_{n,1}^{(2)} + 2\widetilde{T}_{n,1}^{(3)}$ by recalling that $Y_i - m_{\boldsymbol{\theta}_0}(\mathbf{X}_i) = \sigma(\mathbf{X}_i)\varepsilon_i + c_n g(\mathbf{X}_i)$ by (3) and H_{1P}. Specifically, under H_{1P} the conditional variance can be expressed as $\sigma^2(\mathbf{x}) = \mathbb{E}\left[\left(Y - m_{\boldsymbol{\theta}_0}(\mathbf{X}) - c_n g(\mathbf{X})\right)^2 | \mathbf{X} = \mathbf{x}\right] = \sigma^2_{\boldsymbol{\theta}_0}(\mathbf{x})(1 + \boldsymbol{o}(1))$, uniformly in $\mathbf{x} \in \Omega_q$ since g and $\sigma_{\boldsymbol{\theta}_0}$ are continuous and bounded by **A2** and **A7**. Therefore:

$$\widetilde{T}_{n,1}^{(1)} = \int_{\Omega_q} \left(\sum_{i=1}^n \tilde{L}_h(\mathbf{x}, \mathbf{X}_i)\, \sigma(\mathbf{X}_i)\varepsilon_i\right)^2 f(\mathbf{x}) w(\mathbf{x})\, \omega_q(d\mathbf{x}),$$
$$\widetilde{T}_{n,1}^{(2)} = c_n^2 \int_{\Omega_q} \left(\sum_{i=1}^n \tilde{L}_h(\mathbf{x}, \mathbf{X}_i)\, g(\mathbf{X}_i)\right)^2 f(\mathbf{x}) w(\mathbf{x})\, \omega_q(d\mathbf{x}),$$
$$\widetilde{T}_{n,1}^{(3)} = c_n \int_{\Omega_q} \sum_{i=1}^n \sum_{j=1}^n \tilde{L}_h(\mathbf{x}, \mathbf{X}_i)\, \tilde{L}_h(\mathbf{x}, \mathbf{X}_j)\, \sigma(\mathbf{X}_i)\varepsilon_i g(\mathbf{X}_j) f(\mathbf{x}) w(\mathbf{x})\, \omega_q(d\mathbf{x}).$$

By results *ii* and *iii* of Lemma 6, the behavior of the two last terms is

$$nh^{\frac{q}{2}} \widetilde{T}_{n,1}^{(2)} = nh^{\frac{q}{2}} c_n^2 \int_{\Omega_q} g(\mathbf{x})^2 f(\mathbf{x}) w(\mathbf{x})\, \omega_q(d\mathbf{x})(1 + \boldsymbol{o}_{\mathbb{P}}(1)) \text{ and } nh^{\frac{q}{2}} \widetilde{T}_{n,1}^{(3)} = \boldsymbol{o}_{\mathbb{P}}(1). \tag{6}$$

If $c_n^2 nh^{\frac{q}{2}} \to \infty$, then $nh^{\frac{q}{2}}\widetilde{T}_{n,1}^{(2)} \to \infty$, yielding a degenerate asymptotic distribution. If $c_n^2 nh^{\frac{q}{2}} \to 0$, then $nh^{\frac{q}{2}}\widetilde{T}_{n,1}^{(2)} = o_{\mathbb{P}}(1)$. For these reasons, $c_n = \left(nh^{\frac{q}{2}}\right)^{-\frac{1}{2}}$ is assumed from now on. For the first addend, let consider

$$\widetilde{T}_{n,1}^{(1)} = \int_{\Omega_q} \sum_{i=1}^{n} \left(\tilde{L}_h(\mathbf{x}, \mathbf{X}_i)\, \sigma(\mathbf{X}_i)\varepsilon_i\right)^2 f(\mathbf{x}) w(\mathbf{x})\, \omega_q(d\mathbf{x}) + \int_{\Omega_q} \sum_{i \neq j} \tilde{L}_h(\mathbf{x}, \mathbf{X}_i)\, \tilde{L}_h(\mathbf{x}, \mathbf{X}_j)\, \sigma(\mathbf{X}_i)\sigma(\mathbf{X}_j)\varepsilon_i\varepsilon_j f(\mathbf{x}) w(\mathbf{x})\, \omega_q(d\mathbf{x}) = \widetilde{T}_{n,1}^{(1a)} + \widetilde{T}_{n,1}^{(1b)}.$$

From result *iv* of Lemma 6 and because $\sigma^2(\mathbf{x}) = \sigma_{\boldsymbol{\theta}_0}^2(\mathbf{x})(1 + o(1))$ uniformly,

$$nh^{\frac{q}{2}}\widetilde{T}_{n,1}^{(1a)} = \frac{\lambda_q(L^2)\lambda_q(L)^{-2}}{h^{\frac{q}{2}}} \int_{\Omega_q} \sigma_{\boldsymbol{\theta}_0}^2(\mathbf{x}) w(\mathbf{x})\, \omega_q(d\mathbf{x})(1 + o(1)) + o_{\mathbb{P}}(1).$$

The asymptotics of $\widetilde{T}_{n,1}^{(1b)}$ are obtained checking the conditions of Theorem 2.1 in de Jong (1987): *a)* $\mathbb{E}[W_{ijn} + W_{jin}|X_i] = 0$, $1 \leq i < j \leq n$; *b)* $\mathbb{V}\mathrm{ar}[W_n] \to v^2$; *c)* $\left(\max_{1 \leq i \leq n} \sum_{j=1}^{n} \mathbb{V}\mathrm{ar}[W_{ijn}]\right) v^{-2} \to 0$; *d)* $\mathbb{E}[W_n^4]\, v^{-4} \to 3$. To that end, let denote

$$W_{ijn} = \delta_{i,j} nh^{\frac{q}{2}} \int_{\Omega_q} \tilde{L}_h(\mathbf{x}, \mathbf{X}_i)\, \tilde{L}_h(\mathbf{x}, \mathbf{X}_j)\, \sigma(\mathbf{X}_i)\sigma(\mathbf{X}_j)\varepsilon_i\varepsilon_j f(\mathbf{x}) w(\mathbf{x})\, \omega_q(d\mathbf{x}).$$

Then, $nh^{\frac{q}{2}}\widetilde{T}_{n,1}^{(1b)} = W_n = \sum_{i \neq j} W_{ijn}$ and the rv's on which W_{ijn} depends are $(\mathbf{X}_i, \varepsilon_i)$ and $(\mathbf{X}_j, \varepsilon_j)$. *a)* is easily seen to hold by $\mathbb{E}[\varepsilon|\mathbf{X}] = 0$ and the tower property, which implies that $\mathbb{E}[W_{ijn}] = 0$. Because of this, the fact that $W_{ijn} = W_{jin}$ and Lemma 2.1 in de Jong (1987),

$$\mathbb{V}\mathrm{ar}[W_n] = \mathbb{E}\left[\left(\sum_{i \neq j} W_{ijn}\right)^2\right] = 2\mathbb{E}\left[\sum_{i \neq j} W_{ijn}^2\right] = 2n(n-1)\mathbb{E}\left[W_{ijn}^2\right]. \tag{7}$$

Then, by v in Lemma 6 and the fact that $\sigma^2(\mathbf{x}) = \sigma_{\boldsymbol{\theta}_0}^2(\mathbf{x})(1+o(1))$, $\mathbb{E}\left[W_{ijn}^2\right] = n^{-2}\nu_{\boldsymbol{\theta}_0}^2(1 + o(1))$ and as a consequence $\mathbb{V}\mathrm{ar}[W_n] \to 2\nu_{\boldsymbol{\theta}_0}^2$. Condition *c)* follows easily:

$$\left(\max_{1 \leq i \leq n} \sum_{j=1}^{n} \mathbb{V}\mathrm{ar}[W_{ijn}]\right) v^{-2} \leq \left(\max_{1 \leq i \leq n} n^{-1}\nu_{\boldsymbol{\theta}_0}^2(1 + o(1))\right)(2\nu_{\boldsymbol{\theta}_0}^2)^{-1} = (2n)^{-1}(1 + o(1)) \to 0.$$

To check *d)*, note that $\mathbb{E}[W_n^4]$ can be split in the following form in virtue of Lemma 2.1 in de Jong (1987), as Härdle and Mammen (1993) stated:

$$\mathbb{E}\left[W_n^4\right] = 8\sum_{i,j}^{\neq} \mathbb{E}\left[W_{ijn}^4\right] + 12\sum_{i,j,k,l}^{\neq} \mathbb{E}\left[W_{ijn}^2 W_{kln}^2\right] + 48\sum_{i,j,k}^{\neq} \mathbb{E}\left[W_{ijn} W_{ikn}^2 W_{jkn}\right]$$

$$+ 192 \sum_{i,j,k,l}^{\neq} \mathbb{E}\left[W_{ijn}W_{jkn}W_{kln}W_{lin}\right], \tag{8}$$

where $\sum^{\neq}$ stands for the summation over all *pairwise different* indexes (*i.e.*, such that $i \neq j$ for their associated W_{ijn}). By v of Lemma 6, $\mathbb{E}\big[W_{ijn}^4\big] = \mathcal{O}\left((n^4h^q)^{-1}\right)$, $\mathbb{E}\left[W_{ijn}W_{jkn}W_{kln}W_{lin}\right] = \mathcal{O}\left(n^{-4}h^{2q}\right)$ and $\mathbb{E}\left[W_{ijn}W_{ikn}^2W_{jkn}\right] = \mathcal{O}\left(n^{-4}\right)$. Therefore, by (7) and (8),

$$\mathbb{E}\left[W_n^4\right] = 12\sum_{i\neq j}\sum_{k\neq l}\mathbb{E}\left[W_{ijn}^2W_{kln}^2\right] + \boldsymbol{o}\left(1\right) = 3\Big(2\sum_{i\neq j}\mathbb{E}\left[W_{ijn}^2\right]\Big)^2 + \boldsymbol{o}\left(1\right) = 3\mathbb{V}\mathrm{ar}\left[W_n\right]^2 + \boldsymbol{o}\left(1\right)$$

and by **A4**, $\mathbb{E}\left[W_n^4\right] = 3\mathbb{V}\mathrm{ar}\left[W_n\right]^2 + \boldsymbol{o}\left(1\right)$, so *d)* is satisfied, having that

$$nh^{\frac{q}{2}}\widetilde{T}_{n,1}^{(1b)} \xrightarrow{d} \mathcal{N}\left(0, 2\nu_{\boldsymbol{\theta}_0}^2\right). \tag{9}$$

Using the decomposition for T_n with the dominant terms $\widetilde{T}_{n,1}^{(1a)}$, $\widetilde{T}_{n,1}^{(1b)}$ and $\widetilde{T}_{n,1}^{(2)}$, it holds

$$nh^{\frac{q}{2}}T_n = \left(\frac{\lambda_q(L^2)\lambda_q(L)^{-2}}{h^{\frac{q}{2}}}\int_{\Omega_q}\sigma_{\boldsymbol{\theta}_0}^2(\mathbf{x})w(\mathbf{x})\,\omega_q(d\mathbf{x}) + nh^{\frac{q}{2}}\widetilde{T}_{n,1}^{(1b)} + \int_{\Omega_q}g(\mathbf{x})^2f(\mathbf{x})w(\mathbf{x})\,\omega_q(d\mathbf{x})\right)(1+\boldsymbol{o}_{\mathbb{P}}\left(1\right))$$

and the limit distribution follows by Slutsky's theorem and (9). □

Proof of Theorem 2. Analogously as in Theorem 1, $T_n^* = T_{n,1}^* + T_{n,2}^* - 2T_{n,3}^*$.
Terms $T_{n,2}^*$ and $T_{n,3}^*$. By **A9** it is seen that $nh^{\frac{q}{2}}T_{n,2}^* \xrightarrow{p^*} 0$ and $nh^{\frac{q}{2}}T_{n,3}^* \xrightarrow{p^*} 0$, where the convergence is stated in the probability law $\mathbb{P}^*$ that is conditional on the sample.
Term $T_{n,1}^*$. By $\hat{\varepsilon}_iV_i^* = (Y_i - m_{\hat{\boldsymbol{\theta}}}(\mathbf{X}_i))V_i^*$ the dominant term can be split into

$$\begin{aligned}T_{n,1}^* = &\int_{\Omega_q}\sum_{i=1}^{n}\left(W_n^p\left(\mathbf{x},\mathbf{X}_i\right)\hat{\varepsilon}_iV_i^*\right)^2\hat{f}_h(\mathbf{x})w(\mathbf{x})\,\omega_q(d\mathbf{x})\\ &+\int_{\Omega_q}\sum_{i\neq j}W_n^p\left(\mathbf{x},\mathbf{X}_i\right)W_n^p\left(\mathbf{x},\mathbf{X}_j\right)\hat{\varepsilon}_iV_i^*\hat{\varepsilon}_jV_j^*\hat{f}_h(\mathbf{x})w(\mathbf{x})\,\omega_q(d\mathbf{x}) = T_{n,1}^{*(1)} + T_{n,1}^{*(2)}.\end{aligned}$$

From result i of Lemma 7, the first term is

$$nh^{\frac{q}{2}}T_{n,1}^{*(1)} = \frac{\lambda_q(L^2)\lambda_q(L)^{-2}}{h^{\frac{q}{2}}}\int_{\Omega_q}\sigma_{\boldsymbol{\theta}_1}^2(\mathbf{x})w(\mathbf{x})\,\omega_q(d\mathbf{x})(1+\boldsymbol{o}_{\mathbb{P}}\left(1\right)) + \boldsymbol{o}_{\mathbb{P}^*}(1), \tag{10}$$

so the dominant term is $T_{n,1}^{*(2)}$, whose asymptotic behavior is obtained using Theorem 2.1 in de Jong (1987) conditionally on the sample. For that aim, let denote

$$W_{ijn}^* = \delta_{i,j}nh^{\frac{q}{2}}\int_{\Omega_q}W_n^p\left(\mathbf{x},\mathbf{X}_i\right)W_n^p\left(\mathbf{x},\mathbf{X}_j\right)\hat{\varepsilon}_iV_i^*\hat{\varepsilon}_jV_j^*\hat{f}_h(\mathbf{x})w(\mathbf{x})\,\omega_q(d\mathbf{x}).$$

Then, $nh^{\frac{q}{2}}T_{n,1}^{*(2)} = W_n^* = \sum_{i\neq j} W_{ijn}^*$ and the rv's on which W_{ijn}^* depends are now V_i^* and V_j^*. Condition *a)* follows immediately by the properties of the V_i^*'s: $\mathbb{E}^*\left[W_{ijn}^* + W_{jin}^* | V_i^*\right] = 0$. On the other hand, analogously to (7),

$$\mathbb{V}\text{ar}^*\left[W_n^*\right] = 2\sum_{i\neq j}\mathbb{E}^*\left[W_{ijn}^{*2}\right] = 2n^2h^q\sum_{i\neq j}\left[\int_{\Omega_q} W_n^p\left(\mathbf{x},\mathbf{X}_i\right)W_n^p\left(\mathbf{x},\mathbf{X}_j\right)\hat{\varepsilon}_i\hat{\varepsilon}_j\hat{f}_h(\mathbf{x})w(\mathbf{x})\,\omega_q(d\mathbf{x})\right]^2$$

and by result *ii* of Lemma 7, $\mathbb{V}\text{ar}^*\left[W_n^*\right] \xrightarrow{p} 2\nu_{\boldsymbol{\theta}_1}^2$, resulting in the verification of *c)* in probability. Condition *d)* is checked using the same decomposition for $\mathbb{E}^*\left[W_n^{*4}\right]$ and the results collected in *ii* of Lemma 7. Hence $\mathbb{E}^*\left[W_n^{*4}\right] = 3\mathbb{V}\text{ar}^*\left[W_n^*\right]^2 + o_{\mathbb{P}}(1)$ and *d)* is satisfied in probability, from which it follows that, conditionally on $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ the pdf of $nh^{\frac{q}{2}}T_{n,1}^{*(2)}$ converges in probability to the pdf of $\mathcal{N}(0, 2\nu_{\boldsymbol{\theta}_1}^2)$, that is:

$$nh^{\frac{q}{2}}T_{n,1}^{*(2)} \xrightarrow{d} \mathcal{N}\left(0, 2\nu_{\boldsymbol{\theta}_1}^2\right) \text{ in probability.} \tag{11}$$

Using the decomposition of T_n^*, (11) and applying Slutsky's theorem:

$$nh^{\frac{q}{2}}T_n^* = \left(\frac{\lambda_q(L^2)\lambda_q(L)^{-2}}{h^{\frac{q}{2}}}\int_{\Omega_q}\sigma_{\boldsymbol{\theta}_1}^2(\mathbf{x})w(\mathbf{x})\,\omega_q(d\mathbf{x}) + nh^{\frac{q}{2}}T_{n,1}^{*(2)}\right)(1 + o_{\mathbb{P}}(1)) + o_{\mathbb{P}^*}(1).$$

□

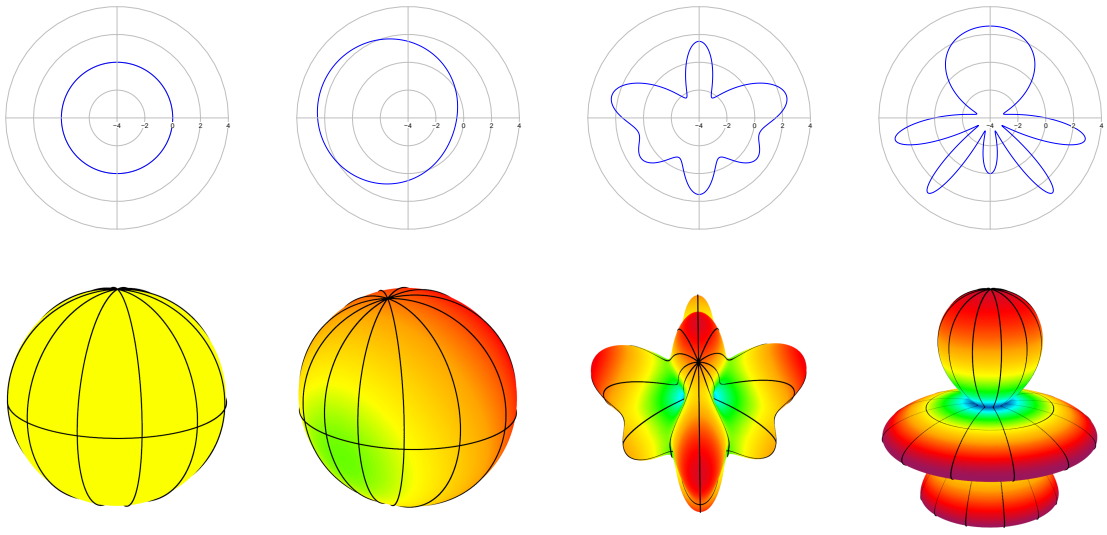

Figure 1: From left to right: parametric regression models for scenarios S1 to S4, for circular and spherical cases. Color shading represents the distance from the origin of the regression surface.

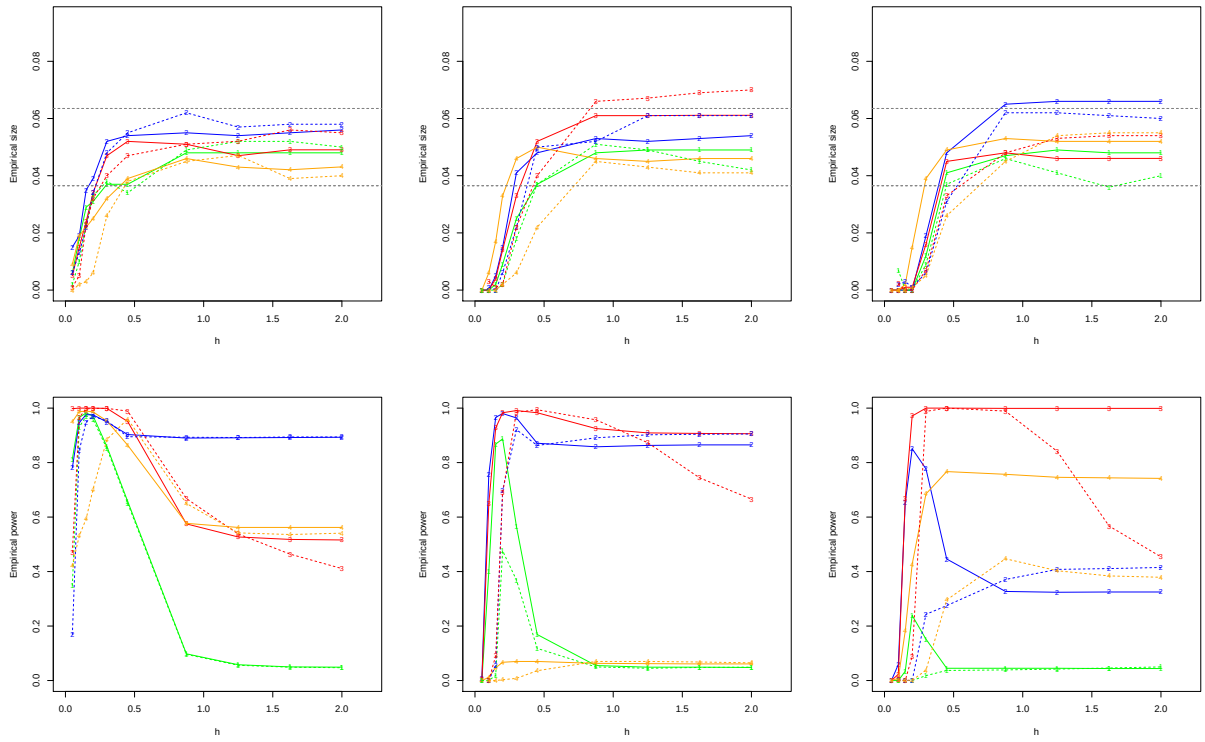

Figure 2: Empirical sizes (first row) and powers (second row) for significance level $\alpha = 0.05$ for the different scenarios, with $p = 0$ (solid line) and $p = 1$ (dashed line). From left to right: columns represent dimensions $q = 1, 2, 3$ with sample size $n = 100$. Green, blue, red and orange colors correspond to scenarios S1 to S4, respectively.

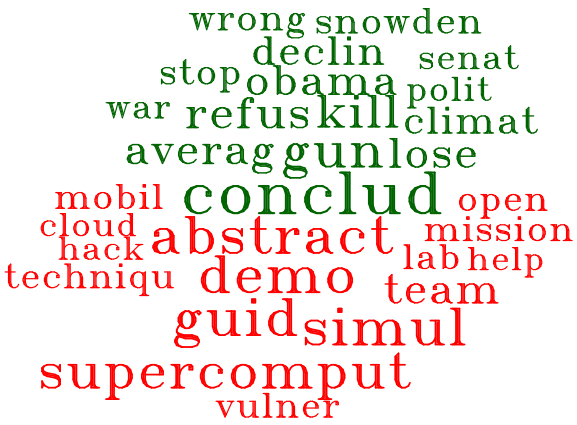

Figure 3: Stems of the 30 largest coefficients (in absolute value) of the fitted constrained linear model. Green and red colors account for positive and negative impacts on news popularity, respectively, whereas the size of the stem is proportional to the magnitude of its coefficient. The linear model has an $R^2 = 0.25$ and the significances of each coefficient are lower than 0.002.